3-22-2013

DQC Dimensions Table Draft 1: An inventory of potential clinical data quality dimensions

An inventory of potential clinical data quality dimensions

Members of the EDM Forum Data Quality Collaborative[1]

Draft: 2012-12-03

A number of recent publications have adapted previous models or developed new models to describe clinical data quality in the context of clinical trials or clinical research. The table below is a synopsis of wide range of clinical data quality dimensions along with potential quantitative data quality assessment methods. We would appreciate your input into developing a consensus set of **core** data quality definitions, operational measures of these definitions, and recommendations for reporting the results of data quality measures in a human- and computer-interpretable format.

Term	**Definition / Possible assessment methods / Examples**
1. Accuracy	Data values represent the true state of a patient within the limitations of the measurement methods. • If gold standard exists, %errors • Units of measure • Rounding rules • Date/time precision
2. Completeness / Missingness	Absence of a value when a data value is expected • Number of NULL (missing) values. • Use of default values to denote missingness (9999, 1/1/1900)
3. Reason missing	The reason that an expected data value is missing coded according to ISO 21090 data type null flavors
4. Contemporaneity	Extent to which a data value is recorded when it is observed. • Date/time recorded - Date/time observed • Date/time observed - Date/time occurred
5. Original	Indication of whether the data value is the initial recording or a certified copy. In the case of an electronic initial recording, a certified copy would include a data transformation using validated code.
6. Legible / relevant	• DDMMYYYY versus MMDDYYYY versus YYYYMMDD • SSN: NNN-NN-NNNN • Mini-sentinel Level 1: name, length, format, acceptable values

[1] In alphabetical order by last name: Jeffrey Brown – HarvardPilgram; Michael Kahn – University of Colorado; Daniela Meeker – Rand Corporation; Meredith Nahm – Duke University; Patrick Ryan – OMOP/J&J; Lisa Schilling – University of Colorado; Nicole Weiskopf – Columbia University; Andrew Williams – Kaiser Permanente Hawaii.

Term	Definition / Possible assessment methods / Examples
	• Conforms to data model • Conforms to terminology/valid elements in value sets
7. Attributable	Data provenance information including data source and individual documenting the data. • Number of variables that can be assigned source and identity of individuals who created or changed data
8. Objectivity	The methods used to obtain data values are well described and represent best practices. Component values represent the total clinical measurement. • Basic aggregate statistics (counts, means, medians, minimum and maximum values • Examination of highest and lowest values • Value distributions (histograms)
9. Believability / Plausibility	Independent measurements are physically plausible and make clinical sense. • Gender = 'M' or 'F' or 'U' only • Route of administration = 'PO', 'IM", 'IV', 'other' only • Blood glucose cannot be a negative value • Age cannot be > 120 years
10. Timeliness / Currency	Serial measurements over time sufficient to detect clinical state
11. Granularity / Appropriate amount of data	Data are present or absent as expected • Number of significant digits or sufficient categories to meet intended use
12. Identity	• Different unique IDs (keys) refer to distinct things (person, place, concept or event) • The same unique ID (e.g. SSN) refers to same entity (e.g. person)
13. Reference	• A reference in one table to data in another table points to a row that exists in the second table.
14. Cardinality	• Relationship cardinality profiling – the count of the actual number of occurrences for each relationship in the database. • References to data in a table refer to no more than the allowed number of occurrences
15. Inheritance	• Entities are grouped into types and subtypes correctly (e.g., all patients, parents, spouses, and employees are also persons)
16. Currency	• The effective date for the earliest record meets a pre-established date. • The date for the most recent record meets a pre-established date.
17. Retention	• The overall duration or number of records per case meets a pre-established threshold (time or number of records).
18. Granularity	• Measures across time all have the same units or duration (e.g. months, years)
19. Continuity	• Gaps or overlaps between records do not exceed pre-specified thresholds.
20. Timeline patterns	• Timestamps fall into expected intervals (e.g. weekly, once a month) • Intervals between successive timestamps do not exceed a minimum or maximum duration.
21. Value patterns	• Successive values follow the expected direction of change (increase or decrease). • Size of change in successive values per unit time is reasonable
22. Event dependencies	• The frequency of events per unit of time does not exceed a pre-specified threshold. • The frequency of events per unit of time meets context-sensitive threshold (e.g., elderly diabetics see their physicians more frequently than healthy teenagers)
23. Event conditions	• One or more events (causes) exist for each observed effect.

Term	Definition / Possible assessment methods / Examples
	• Certain events always occur together (coincidental events)
24. Event attributes	• All attributes relevant to an event description are present (e.g., a surgical event requires a patient and a surgeon).
25. State-transition profiling	• Examination of valid transition states over time. • A terminal state cannot be followed by another state (e.g., an expired patient cannot be readmitted). • A valid action or event is associated with a corresponding state transition (e.g., cardiac arrest precedes patient state transition to expired).
26. State domain	• An object's state can only be a valid value (e.g. either admitted or discharged from hospital).
27. Action domain	• The set of actions that can be applied to an object can only be a valid value (e.g. a test can be performed or cancelled, but not both).
28. Terminator domain	• States in which an object can start or stop its lifecycle can only be a valid value.
29. State-actions	• Each action is consistent with the change in state that it engenders (e.g., myocardial infarction cannot lead to subsequent "no cardiac disease" state).
30. Continuity	• Sequence or timing of start of each state record must follow the end of the previous state record.
31. Duration	• The minimum or maximum length of time an object can stay in a specific state (e.g., admission for myocardial infarction cannot be less than 1 day). • Zero-length rule: End date of a state must be later than the start date.
32. Redundant attributes	• Same attributes in different data sources should have identical values • Historical measures of a non-time varying attribute should have identical values
33. Derived attributes	• An aggregated value must equal the total of the atomic level values. • An aggregated value must follow appropriate rules when component data elements are missing
34. Partially dependent attributes	• The allowed values of one attribute are limited by the assigned value of another attribute (e.g., sex = female eliminates prostate cancer as a valid diagnosis).
35. Conditional optionality	• The allowed values of one attribute determines if another attribute must be (cannot be) present (e.g., a discharge disposition = "to home" implies expiration date = null). • Mutually exclusivity – the presence of any value in one attribute precludes (requires) a value in another attribute
36. Correlated attributes / Concordance	• Values in one attribute changes the likelihood of values in another attribute (e.g., sex = male and age = 65 and smoker = yes increases the likelihood of discharge diagnosis = myocardial infarction) • Mini-sentinel Level 2: integrity between variables within a table or variables across tables.

Sources: Brown: (under review); Hersh: (under review); Kahn: Medical Care 2012, Nahm: Chapter in Richesson book; Weiskopf: JAMIA 2012.
Some terms modified from Maydanchik, Wang & Strong.